Archives of Toxicology
https://doi.org/10.1007/s00204-019-02520-8

IN VITRO SYSTEMS

Model-based estimation of lowest observed effect concentration from replicate experiments to identify potential biomarkers of in vitro neurotoxicity

Silvia Calderazzo[1] · Denise Tavel[2,3] · Marie-Gabrielle Zurich[2,3] · Annette Kopp-Schneider[1]

Received: 29 March 2019 / Accepted: 10 July 2019

Abstract

A paradigm shift is occurring in toxicology following the report of the National Research Council of the USA National Academies entitled "Toxicity testing in the 21st Century: a vision and strategy". This new vision encourages the use of in vitro and in silico models for toxicity testing. In the goal to identify new reliable markers of toxicity, the responsiveness of different genes to various drugs (amiodarone: 0.312–2.5 µM; cyclosporine A: 0.25–2 µM; chlorpromazine: 0.625–10 µM; diazepam: 1–8 µM; carbamazepine: 6.25–50 µM) is studied in 3D aggregate brain cell cultures. Genes' responsiveness is quantified and ranked according to the Lowest Observed Effect Concentration (LOEC), which is estimated by reverse regression under a log-logistic model assumption. In contrast to approaches where LOEC is identified by the first observed concentration level at which the response is significantly different from a control, the model-based approach allows a principled estimation of the LOEC and of its uncertainty. The Box–Cox transform both sides approach is adopted to deal with heteroscedastic and/or non-normal residuals, while estimates from repeated experiments are summarized by a meta-analytic approach. Different inferential procedures to estimate the Box–Cox coefficient, and to obtain confidence intervals for the log-logistic curve parameters and the LOEC, are explored. A simulation study is performed to compare coverage properties and estimation errors for each approach. Application to the toxicological data identifies the genes *Cort*, *Bdnf*, and *Nov* as good candidates for in vitro biomarkers of toxicity.

Keywords Dose–response modeling · Log-logistic model · Box–Cox transform both sides · 3D cultures · Neurotoxicity

Introduction

There is nowadays a large consensus that for toxicology animal testing needs to be replaced by a combination of in silico and in vitro approaches, as evidenced by ethical and economic arguments, besides scientific ones. Clearly (neuro)toxicity testing will soon be based on technologies leading to the better understanding of biological processes to identify the effects of chemicals on toxicity pathways. Such an approach is discussed in the report of the National Research Council of the USA National Academies entitled "Toxicity testing in the 21st Century: a vision and strategy" (National Research Council 2007). This report proposes to focus testing at the molecular level of toxicity pathways using an in vitro approach rather than observing phenotypic responses at the level of whole organisms as it has been done so far. Additionally, the interpretation of chemically induced alterations in toxicity pathways will be based on sophisticated modeling that will extrapolate

Marie-Gabrielle Zurich and Annette Kopp-Schneider shared the senior authorship.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00204-019-02520-8) contains supplementary material, which is available to authorized users.

✉ Silvia Calderazzo
s.calderazzo@dkfz.de

Marie-Gabrielle Zurich
mzurich@unil.ch

Annette Kopp-Schneider
kopp@dkfz.de

[1] Division of Biostatistics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

[2] Department of Physiology, University of Lausanne, Rue du Bugnon 7, 1005 Lausanne, Switzerland

[3] Swiss Center for Applied Human Toxicology (SCAHT), University of Lausanne, Lausanne, Switzerland

Published online: 19 July 2019

from measured dose–response relationships in cell-based systems to human exposure (Bal-Price et al. 2010).

Aggregating brain cell cultures were extensively used for neurotoxicological investigations and proved to be suitable for the detection of organ-specific toxicity (Zurich et al. 2013). These 3D cultures of rat brain cells are composed of neurons, astrocytes, oligodendrocytes and microglial cells. The model allows multiple cell–cell interactions, and the development of histotypic structures such as extracellular matrix, synapses and myelinated axons (Honegger et al. 1979). Several structural and functional endpoints were shown to be useful specific markers of neurotoxicity, such as the activity of cell type-specific enzymes, the expression of selected genes, as well as astroglial and microglial reactivities (Monnet-Tschudi et al. 1995; Zurich et al. 2004, 2013). In the goal to improve the predictive capacity of in vitro toxicity testing to alert for neurospecific toxicity, we have looked for new early markers of adverse effects. Concentration–response experiments are performed in which 3D rat brain cell cultures are exposed to several concentrations of various drugs and the expression of numerous genes is measured. The aim is to rank the genes according to their responsiveness to the compounds. Rather than identifying the lowest tested concentration that is statistically significant from control, a model-based approach is chosen to estimate the Lowest Observed Effect Concentration (LOEC) to be independent of the actually tested concentration levels. Hence, the main focus of the data analysis is to use reverse regression to estimate the concentration at which a prespecified drop or increase in response can be inferred for every chemical and every gene, and to rank the genes according to the size of the estimated LOEC. This toxicological question provides the motivation for a statistical investigation of different approaches for LOEC estimation from replicate experiments using nonlinear dose–response modeling. Here we focus on a log-logistic dose–response model, which is a common choice in this modeling framework (see, e.g., Ritz and Streibig 2008).

A difficulty which frequently arises in nonlinear dose–response modeling is that residuals may be characterized by heteroscedasticity and/or non-normality. Different solutions have been proposed in the literature to address this issue. Heteroscedasticity can be addressed by further modeling of the variance, while the impact of non-normality and/or heteroscedasticity may be improved by transformation or robust estimation (Ritz and Streibig 2008). The Box–Cox transform both sides approach, introduced by Carroll and Ruppert (1984), is a transformation approach which has the advantage of preserving parameter interpretability by transforming both the response and the mean function. The methodology has been applied by Ritz and Van der Vliet (2009) in the context of toxicological studies.

The modeling approach to LOEC estimation is a means to provide a confidence interval for the estimate. The Delta method represents a standard approach to reach this purpose and is based on Wald confidence intervals for the model parameters, i.e. both the parameters and the LOEC estimates are assumed to be approximately normally distributed. However, the non-linearity of both the mean function and of the parameter coordinates may affect the precision of Wald confidence intervals (Bates and Watts 1988; Ritz and Streibig 2008). This leads to the second difficulty to be addressed in the modeling framework. Non-linearity in the parameter coordinates may be tackled by resorting to likelihood-based confidence intervals (Carroll and Ruppert 1988; Ritz and Streibig 2008) or to bootstrap (Ritz and Streibig 2008).

A third challenge is posed by the availability of replicate experiments. Replicate experiments allow to disentangle intra- and inter-experimental variability, and can strengthen estimation of the single-experiment parameters through borrowing of information. Mixed effects models and meta-analytic approaches can be adopted to perform estimation of the single-experiment parameters and of their global mean and variance (see Jiang and Kopp-Schneider 2014, 2015, and references therein).

We approach estimation of the dose–response model parameters and LOEC adopting the Box–Cox transform both sides approach to improve the distributional properties of the residuals. The Box–Cox transform both sides is a flexible approach which is not restricted to a specific distributional assumption for the residuals, and includes the log-normal assumption as a sub-case. We perform a simulation study comparing relative strengths and weaknesses of different approaches on estimation of the Box–Cox parameter, the log-logistic function parameters, and the LOEC. We additionally compare Wald and bootstrap confidence intervals, and the corresponding summary estimates for multiple experiments provided by the meta-analytic approach. The work of Latif and Gilmour (2015) is closely related to our approach. Our simulation study mainly extends it by including semi-parametric and non-parametric bootstrap confidence intervals, and by focusing on a four-parameter log-logistic function (although only three parameters are effectively estimated) as compared to a two-parameters Michaelis–Menten one; moreover, we focus on the meta-analytic approach to combine estimates from different biological replicates, as compared to a random effects model. We focus on generally smaller sample sizes and number of replicates than Latif and Gilmour (2015), and the meta-analytic approach is indeed an attempt to cope with estimation when only three biological replicates are available, a situation which is not uncommon in toxicological studies.

In the next section, we introduce the model, the inferential approaches, and the available data. The simulation study is presented in the third section. Finally, the application of the

model to the available data and the relevant toxicological results are illustrated in the fourth section. In the last section, the results of simulation study and real data application are discussed and a recommendation is given.

Methods

Model

Given a set of doses $x = \{x_1, \ldots, x_n\}$ corresponding to a sample of response measurements $y = \{y_1, \ldots, y_n\}$, the relationship between x and y is commonly described by a nonlinear model of the form:

$$y_i = g(x_i, \theta) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma), \tag{1}$$

where $g(\cdot)$ is an appropriate nonlinear function describing the mean of y, and θ is a vector of parameters. In this work, we focus on the log-logistic mean function:

$$g(x_i, \theta) = c + \frac{d - c}{1 + \exp\{b(\log(x) - e)\}},$$

where $\theta^{\mathrm{T}} = [b, c, d, e]$ is the parameter vector. The parameter c represents the lower function asymptote, d its upper asymptote, b relates to its steepness, and $\exp(e)$ is commonly known as EC_{50}, i.e., the dose level required to achieve a half-maximal effect in the response. In the present toxicological setting, interest is placed upon the LOEC, determined from inverse regression by:

$$\mathrm{LOEC} = \exp(e)\left(\frac{d - \lambda}{\lambda - c}\right)^{1/b},$$

where λ is the targeted response level. To visually illustrate the concept, Fig. 1 displays the available dataset of a sample drug–gene combination, to which the log-logistic curve fit and estimated LOEC are superimposed. In the current setting, data are normalized to mean control and either the upper (for decreasing) or the lower asymptote (for increasing relationships) is fixed to one. Since a deregulation of gene expression of a factor of 1.7 is to be detected, λ is chosen equal to $1/1.7 = 0.59$ for a decreasing relationship, and 1.7 for an increasing one in accordance with previous evaluations of this type of experimental data (e.g., Zurich et al. 2013).

The model in (1) induces a normal likelihood, which can be maximized to obtain estimates and confidence intervals for the parameters θ and σ.

Under the assumption of approximate normality for the parameter estimates, Wald confidence intervals can be derived. Such intervals are characterized by a symmetric shape, and may suffer by lack of accuracy in such nonlinear model (Bates and Watts 1988; Ritz and Streibig 2008). The Delta-method which is based on the assumption of approximate normality of a transformation of normal variables may be applied to derive confidence intervals for the LOEC.

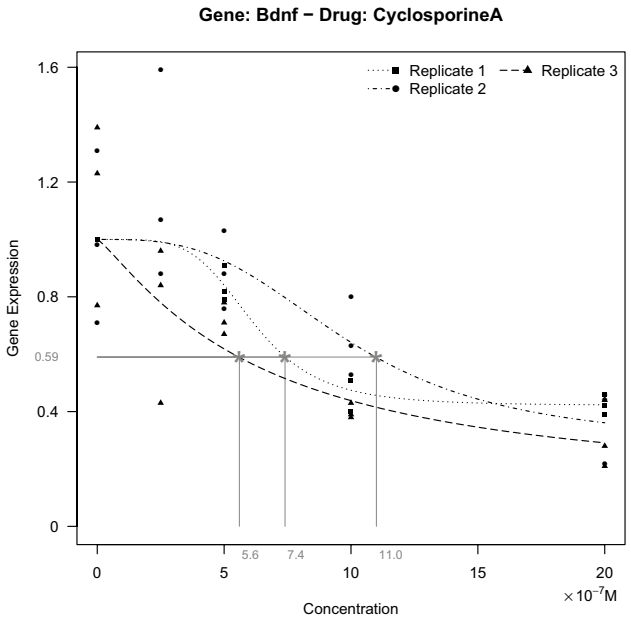

Fig. 1 Sample available toxicological dataset. The log-logistic curve fit and LOEC estimate (for $\lambda = 0.59$) are superimposed for illustration

Bootstrap confidence intervals represent a more accurate and correct alternative to Wald and Delta-method confidence intervals in this scenario (Ritz and Streibig 2008). Such intervals are obtained through resampling of the available data (non-parametric bootstrap), resampling of the model residuals (semi-parametric bootstrap), or by drawing new observations from the model fitted on the available data (parametric bootstrap) (see Carroll and Ruppert 1988, and Section 1.1 of Online Resource 1 for further details on the latter two approaches). Each bootstrap dataset provides an estimate of the parameters and of the LOEC, i.e. samples from their bootstrap distribution. From the latter, confidence intervals may be derived according to different methods (see e.g. Efron and Tibshirani 1994). Bootstrap confidence intervals do not rely on the assumption of approximate normality of the parameter estimates. Nevertheless, their performance is influenced by how well the available data allow estimating the true data-generating process.

Box–Cox transform both sides approach

One difficulty which may arise with model (1) is that response levels are constrained to be positive, which may contribute to induce non-normality in the residuals. Heteroscedasticity may also be present. Transformations can be adopted to improve the distributional properties of the residuals (Ritz and Streibig 2008). A flexible family of

transformations is provided by the Box–Cox family (Box and Cox 1964):

$$h(y) = \begin{cases} \frac{(y+\kappa)^{\phi}-1}{\phi} & \text{for } \phi \neq 0 \\ \log(y+\kappa) & \text{for } \phi = 0 \end{cases}$$

indexed by a parameter $\phi \in \mathbb{R}$, and, if required, κ such that $y + \kappa > 0$. We assume from now on κ to be fixed. The Box–Cox transformation can be applied to both the response and the mean function, (Carroll and Ruppert 1984, 1988; Ritz and Van der Vliet 2009), i.e.,

$$h(y_i) = h(g(x_i, \theta), \phi) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma).$$

Note that, if the Box–Cox transformation h (or in general any monotonic transformation of both the response and the predictor) leads to approximate symmetry of the response $h(Y)$, then the approach is effectively equivalent to performing median regression, and estimation is more efficient than minimizing the least absolute deviations (Fitzmaurice et al. 2007), i.e., minimizing the absolute value of the residuals. Therefore, parameter interpretation (including interpretation of the LOEC) remains the same for any suitable choice of ϕ, even if ϕ differs across different datasets. Moreover, if the main interest lies in prediction, in the mean, or in alternative quantiles of y, the corresponding prediction or confidence intervals can also be easily derived under fulfillment of approximate normality of $h(Y)$ (Carroll and Ruppert 1988).

Other approaches involving transformations of x and/or y according to a chosen performance criterion (additive regression, alternating conditional expectation, and additivity and variance stabilization, see e.g. Harrell 2015), as well as quantile regression, semi-parametric approaches or model averaging, can also handle the violation of model assumptions and allow LOEC estimation. A problem with the above-mentioned transformation approaches is that they may result in over-fitting and, therefore, be poorly generalizable, particularly for small sample sizes (approximately below 100 samples, Harrell 2015), and only relatively small sample sizes are often available in toxicological studies. Quantile regression would fit a parametric dose–response model for, e.g., the $\alpha/2$ and $1-\alpha/2$ quantile, and then inverse regression can be applied to obtain the dose interval inducing the response level of interest at each quantile (see Jensen et al. 2019). A large sample size is, however, generally required also by this approach (Wheeler et al. 2015; Jensen et al. 2019). Semi-parametric approaches (see, e.g., Nottingham and Birch 2000) and model averaging (see, e.g., Ritz et al. 2013) can also provide sensible robust alternatives to the Box–Cox transform both sides approach. The semi-parametric approach proposed by Nottingham and Birch (2000) consists in fitting a weighted combination of a parametric (logistic) model and a non-parametric one (local linear regression). The approach of Ritz et al. (2013) requires the selection and fitting of a set of candidate models, and the estimates of the quantity of interest is then averaged according to weights related to the AIC or BIC information criterion. Robustness comes, however, at the cost of an increased complexity: as noted by the respective authors, the semi-parametric approach requires the appropriate selection of the model mixing and tuning parameters, while model averaging requires the specification of a pool of candidate models. Due to the limited sample size in our real data application, and the fact that we do not encounter strong evidence of deviation from the assumed log-logistic model, nor of residual non-normality after the Box–Cox transformation, we do not pursue alternative approaches.

The likelihood function is obtained after multiplication with the Jacobian of the transformation, and ϕ can be treated as an additional parameter to be estimated. In practice, ϕ is often estimated a priori, and kept constant while performing inference on the remaining parameters. When only the response is transformed, ignoring uncertainty about ϕ can have a strong impact on the confidence intervals of the remaining parameters (Bickel and Doksum 1981). However, in the transform both sides approach, this would impact significantly only on the confidence interval of σ (Carroll and Ruppert 1984, 1988), which is generally not of interest. Two main choices are available to estimate ϕ:

- minimization of the residual sum of squares in an ANOVA analysis (Box and Cox 1964; Carroll and Ruppert 1988; Ritz and Van der Vliet 2009; Latif and Gilmour 2015);
- maximum likelihood estimation (Box and Cox 1964; Carroll and Ruppert 1988; Ritz and Streibig 2008; Latif and Gilmour 2015).

Each approach is applied to a grid of ϕ values, and the (approximately) optimal estimate is identified. The ANOVA approach requires replicates at each dose level. Observations should be rescaled by dividing each transformed observation by the overall geometric mean, to ensure that each value of ϕ induces approximately the same scale on the transformed data (Hinkley and Runger 1984). The maximum likelihood approach is more precise but slower, and relies on nonlinear optimization.

Inference for replicate experiments

If data from experimental replicates are available, a final summary of the results which properly accounts for intra- and inter-experimental variability is sought. Moreover, combining data from replicate experiments in the estimation process can improve inference through borrowing of information.

Random-effect models are commonly used for this purpose, under the assumption that each experiment-specific parameter (we will refer from now on to these parameters as 'level one' parameters) is drawn from a common distribution, whose parameters are in turn estimated (we will refer from now on to these parameters as 'level two' parameters). In the nonlinear framework, this has proven to be challenging and an alternative is provided by a meta-analytic approach, which does not allow to share information across groups, but provides a robust alternative to properly estimate level two parameters (Jiang and Kopp-Schneider 2014, 2015). We focus on the LOEC as our parameter of interest, and, to improve compliance with the usual normal distributional assumption, the logarithmic scale is adopted.

Let k, $k = 1, \ldots, K$, be the experiment index. The meta-analytic approach assumes

$$\begin{aligned}\log(\text{LOEC})_k =& \theta + \epsilon_k \\ \theta \sim& N(\mu_{\log(\text{LOEC})}, \sigma_{\log(\text{LOEC})}) \\ \epsilon_k \sim& N(0, \sigma_{\epsilon_k}),\end{aligned}$$

and generally aims at obtaining an estimate and a confidence interval for the mean effect $\mu_{\log(\text{LOEC})}$, given unpooled estimates of the parameter means and standard deviations. Such estimates can be obtained in a first model fitting of the unpooled data.

Toxicological data

3D rat brain cell cultures are grown in flasks containing between 200 and 500 spheres, as previously described Zurich et al. (2013). Cells are exposed to four concentrations of amiodarone (0.312–2.5 µM), cyclosporine A (0.25–2 µM), diazepam (1 − 8 µM), carbamazepine (6.25–50 µM), and five concentrations of chlorpromazine (0.625–10 µM) for 24 h or 14 days. Control cultures receive the equivalent amount of the solvent used to dissolve the drug (DMSO, final concentration 0.05%). qRT-PCR analyses are performed, as described in (Zurich et al. 2013), to quantify gene expression of glial fibrillary acidic protein (*Gfap*), heat shock protein 32 (*Hsp32*), myelin basic protein (*Mbp*), and neurofilament high molecular weight (*Nfh*), whereas Taqman gene expression assays (Life Technologies) are used for: brain-derived neurotrophic factor (*Bdnf*, Rn02531967_s1), cortistatin (*Cort*, Rn00563272_m1), and nephroblastoma over-expressed gene (*Nov*, Rn00578390_m1). Three independent experiments are performed with each time three replicate cultures per concentration or control. Data are then normalized so that the mean response value at concentration zero is equal to one. Henceforth, we focus on the long-term experiment. Figures 1–3 in Online Resource 2 provide plots of the raw data.

Simulation study

We perform a simulation study to investigate and compare the performance of the different estimation algorithms. Simulated data reproduce the available information, i.e., we assume that three experimental replicates are available, each comprising three replicate observations at five concentration levels. Parameters are chosen to approximately reproduce observed dynamics, and in particular we assume two scenarios, a decreasing and an increasing dose–response relationship. We simulate 100 datasets comprising three experimental replicates for each scenario. For the decreasing dose–response relationship, we draw values for b from a truncated normal distribution $TN(1.2, 1)$, with lower truncation at 0. The constraint of b to positive real values ensures parameter identifiability. The lower asymptote c is drawn from a $TN(0.2, 0.3)$ with lower truncation at 0, and upper truncation at $\lambda = 0.59$. While the lower constraint is motivated by the biological interpretation of the parameter (gene expression cannot be negative), the upper constraint ensures existence of the LOEC. For the increasing dose–response relationship, we draw b from a $TN(-2, 1)$, with upper truncation at 0, and d is drawn from a $TN(2, 0.3)$ with lower truncation at $\lambda = 1.7$. For both scenarios, the Box–Cox parameter ϕ is assumed equal to 0 in all datasets, so that normal observations can be drawn in the transformed scale; κ is assumed equal to 0.05, $\log(\sigma_y)$ is drawn from a $N(0.2, 0.1)$, and log(LOEC) is drawn from a $N(-14, 1)$. The normally distributed responses are then transformed by applying the function h^{-1}.

Prior to model fitting, datasets are filtered according to their mean response among all concentration levels and only samples in which the minimum mean response is below (for a decreasing relationship) or the maximum mean response is above (for an increasing relationship) λ, are retained. This filtering step is performed to mimic the estimation process later applied to the real data, which aims at preselecting data sets for which LOEC is likely to exist and fall within the observed range of concentrations. For the simulated datasets, the filtering process drops 47 and 51 individual experiments datasets (out of 300) for scenario 1 and 2, respectively.

The model is then fitted on the unpooled data via the *drm* function in the `drc` R package (Ritz et al. 2016), in an unconstrained form. Full details are provided in the Section 1.1 of Online Resource 1.

We compare coverage properties and length of the confidence intervals obtained under the following scenarios and their combinations:

- the parameter ϕ is estimated either via (1) ANOVA, (2) maximum likelihood estimation (ML), or (3) no transformation is applied;

- confidence intervals are derived as (1) standard Wald for the model parameters and via the Delta method for log(LOEC), or (2) nonparametric bootstrap, (3) semi-parametric bootstrap, (4) parametric bootstrap;
- starting points for optimization are (1) unique, provided by a self-starter function or (2) multiple, provided by random perturbation of the self-starter function values.

Results for the LOEC are shown in Table 1. Focusing first on the single-experiment results, it is somewhat surprising that the Wald-Delta method confidence intervals always behave equally or better than bootstrap confidence intervals in terms of accuracy, with coverage probabilities in the range of 87–95% as compared to 72–88% for the bootstrap intervals, although at the cost of an increased length (median length in the range 1.1–1.7 as compared to 0.8–1.5). The result may be due to a poor estimation of the cumulative distribution function given the small sample size, especially in the tails and/or possible convergence failures of the estimation algorithm. The Wald-Delta method confidence intervals seem to counterbalance such problems thanks to their increased width. Adopting perturbed starting points seems overall not to provide a consistent advantage in terms of coverage; median squared error values (see Table 3 in Online Resource 2) seem to be slightly decreased when adopting perturbed initial conditions. Finally, applying the mean and data transformation seems to generally either reduce the length of the intervals at the cost of a lower coverage (scenario 1), or otherwise improve coverage at the cost of longer intervals (scenario 2). The result is again probably influenced by occasional non-convergences of the algorithm. No additional patterns are observed across the two simulation scenarios, except a possible superiority of the semi-parametric bootstrap in terms of coverage, if compared with its non-parametric and parametric counterparts.

Table 1 Coverage probability (CP), median confidence interval length (MIL) and number of single-experiment level converged estimates of log(LOEC) (n) and of $\mu_{\log(LOEC)}$ (n_H), for model fitting of dose–response data from two simulation scenarios: a decreasing (scenario 1) and an increasing (scenario 2) relationship

Approach	Scenario 1 ($b > 0$)						Scenario 2 ($b < 0$)					
	log(LOEC)			$\mu_{\log(LOEC)}$			log(LOEC)			$\mu_{\log(LOEC)}$		
	CP	MIL	n	CP	MIL	n_H	CP	MIL	n	CP	MIL	n_H
Self-starter starting points												
ANOVA Wald-Delta	0.90	1.25	230	0.88	2.57	48	0.90	1.65	218	0.71	2.12	38
ANOVA bootstrap semi-par.	0.83	1.08	239	0.72	1.56	54	0.78	1.52	228	0.61	1.54	44
ANOVA bootstrap par.	0.75	0.94	239	0.72	1.88	54	0.72	1.24	228	0.66	1.71	44
ANOVA bootstrap non-par.	0.74	0.94	239	0.78	2.48	54	0.74	1.31	226	0.60	1.76	42
ML Wald-Delta	0.88	1.15	231	0.94	2.82	48	0.91	1.61	219	0.72	2.11	39
ML bootstrap semi-par.	0.77	0.99	241	0.75	2.53	55	0.79	1.45	226	0.62	1.55	42
ML bootstrap par.	0.72	0.78	241	0.76	2.49	55	0.73	1.25	226	0.62	1.94	42
ML bootstrap non-par.	0.74	0.94	240	0.82	2.52	55	0.74	1.32	225	0.61	1.44	41
Untr. Wald-Delta	0.95	1.58	228	0.85	2.55	46	0.87	1.40	220	0.78	1.84	37
Untr. bootstrap semi-par.	0.88	1.42	235	0.68	1.19	50	0.76	1.38	238	0.62	1.56	48
Untr. bootstrap par.	0.84	1.32	235	0.62	1.13	50	0.76	1.24	238	0.56	1.36	48
Untr. bootstrap non-par.	0.75	0.92	237	0.85	2.56	52	0.74	1.23	236	0.52	0.91	46
Perturbed starting points												
ANOVA Wald-Delta	0.90	1.26	230	0.88	2.67	48	0.91	1.67	214	0.75	2.11	36
ANOVA bootstrap semi-par.	0.81	1.06	238	0.70	1.50	54	0.77	1.47	226	0.67	1.48	43
ANOVA bootstrap par.	0.75	0.90	238	0.69	1.43	54	0.72	1.27	226	0.70	1.47	43
ANOVA bootstrap non-par.	0.74	0.94	237	0.76	2.02	54	0.73	1.23	226	0.58	1.32	43
ML Wald-Delta	0.89	1.13	229	0.91	2.86	47	0.91	1.65	217	0.73	2.28	37
ML bootstrap semi-par.	0.78	1.04	238	0.80	1.83	54	0.78	1.45	228	0.67	1.74	43
ML bootstrap par.	0.74	0.87	238	0.76	1.85	54	0.74	1.26	228	0.70	1.47	43
ML bootstrap non-par.	0.76	0.98	237	0.78	2.34	54	0.74	1.25	227	0.61	1.41	44
Untr. Wald-Delta	0.94	1.58	227	0.84	2.65	45	0.88	1.40	216	0.81	2.32	37
Untr. bootstrap semi-par.	0.88	1.40	234	0.67	1.06	51	0.77	1.42	237	0.54	1.33	46
Untr. bootstrap par.	0.84	1.28	235	0.56	0.81	52	0.75	1.20	237	0.59	1.36	46
Untr. bootstrap non-par.	0.72	0.93	236	0.81	2.10	53	0.74	1.16	237	0.54	0.93	46

Results obtained for 300 individual experiments (100 datasets), and different combinations of confidence intervals estimation approaches, data transformations and initial conditions for the likelihood optimization. Unconstrained optimization

The meta-analytic approach provides $\mu_{\log(\mathrm{LOEC})}$ estimates and confidence intervals. Coverage is overall reduced (being now between 52 and 94%) and median confidence interval lengths are observed in the range between 0.8 and 2.6. Patterns similar to those encountered in the unpooled analysis are observed for accuracy and interval length.

Coverage probabilities and median confidence interval lengths for the unpooled estimates of the remaining log-logistic function parameters are provided in Tables 1 and 2, respectively, in Online Resource 2. The behavior of the different approaches is similar to that of the LOEC estimates, except for the fact that the median length of the bootstrap confidence intervals is larger than that of the Wald intervals in several cases. The explanation may be found in the skewed bootstrap distributions with extremely long tails which are sometimes observed among the bootstrap samples, possibly caused in turn by a convergence failure of the estimation algorithm. In such cases, extreme values may shift and extend the intervals in the direction of the longest tail, thus failing to include the true value. Perturbed starting points seem also to provide slightly increased median squared error values which suggests the presence of a multimodality in the target log-likelihood.

Table 2 Ranking of the genes according to $\mu_{\log(\mathrm{LOEC})}$, by drug

Gene	Estimate	[95% CI]
Amiodarone		
Cort	− 17.51	[− 19.60, − 15.42]
Bdnf	− 14.97	[− 20.04, − 9.89]
Nov	− 14.40	[− 15.24, − 13.55]
Nfh	− 14.19	[− 14.59, − 13.80]
Hsp32	− 13.75	[− 14.69, − 12.80]
Cyclosporine A		
Nov	− 15.90	[− 17.88, − 13.91]
Nfh	− 14.10	[− 14.45, − 13.75]
Bdnf	− 14.06	[− 14.83, − 13.30]
Chlorpromazine		
Cort	− 13.84	[− 14.92, − 12.76]
Nfh	− 13.71	[− 15.62, − 11.81]
Bdnf	− 13.64	[− 14.28, − 13.01]
Gfap	− 13.35	[− 14.68, − 12.02]
Nov	− 13.05	[− 14.02, − 12.08]
Hsp32	− 12.91	[− 13.23, − 12.58]
Diazepam		
Nov	− 13.53	[− 15.49, − 11.58]
Bdnf	− 13.03	[− 15.88, − 10.19]
Cort	− 12.97	[− 14.12, − 11.82]
Carbamazepine		
Bdnf	− 11.29	[− 12.91, − 9.67]
Cort	− 10.83	[− 11.30, − 10.35]

Maximum likelihood estimate and Knapp and Hartung adjusted 95% confidence intervals (CI) are displayed (the confidence intervals for the unpooled datasets are computed via the Wald-Delta method, with perturbed initial values and an ANOVA estimate of ϕ)

We further explore the effect of constraining the asymptotes to positive values, and the slope to either positive or negative values. As shown in Table 4 in Online Resource 2, coverage probabilities and length of the intervals are barely affected. The most remarkable difference is a decrease in the number of fitted datasets, which we attribute to the additional challenges posed by the boundaries in optimization.

Finally, we run an analogous simulation study for an additional design, performed on a 96-well plate, which represents a common choice in cytotoxicity studies (Anon 2006). The design consists of 8 concentrations with 6 replicates each, plus the control with 12 replicates. Concentrations are selected so that the hypothesized EC_{50} (the concentration closest to it in a preceding range finder test) represents their mid-point, and are spaced, e.g., according to a serial dilution factor of 1.47. We assume an $\mathrm{EC}_{50} = \exp\{-14\} = 8.31e-07$. The remaining simulation parameters are left unchanged. Three biological replicates are also again assumed. Results for this 96-well design are reported in Tables 5–9, in Online Resource 2. Table 5 shows the simulation results for the LOEC estimates: although Wald-Delta method confidence intervals coverage seem to retain some superiority as compared to bootstrap intervals, the difference is less pronounced. The length of the intervals is significantly shorter in this design due to the increased sample size, however, bootstrap confidence intervals may be somewhat wider in some scenarios, possibly again due to occasional non-convergences. Perturbation of the starting points appears not to provide any advantage overall, and occasionally worsen results also in terms of median squared errors (see Table 8). This phenomenon points in the direction of a multi-modal target, whose exploration may become more challenging when the sample size increases. Analogous conclusions can be drawn for the log-logistic model parameters (Tables 6–8). With respect to the effect of adopting a constrained optimization algorithm, Table 9 shows that this approach could lead to relevant advantages for the bootstrap intervals coverage. The reduction in the number of converged estimates is also weaker if compared to the previous design, and affects more significantly scenario 2.

Application to toxicological data

The methodology is finally applied to the available toxicological data. Since data have been normalized so that the mean response value at concentration zero is equal to 1, the upper (for decreasing relationships) or lower (for increasing relationships) asymptote is fixed to this value.

Wald-Delta method confidence intervals are computed for a given ANOVA-based estimate of ϕ and perturbed initial

values. The choice is based on the observed coverages, MSE and interval lengths, as well as the number of converged samples and normality of the residuals.

Table 2 provides ranking of the genes according to the meta-analytic estimate of $\mu_{\log(\mathrm{LOEC})}$, and Knapp and Hartung adjusted 95% confidence intervals are included. When confidence intervals are not available for one or more of the individual-experiment estimates, the result is not summarized in a meta-analytic estimate. All available individual-experiment log(LOEC) estimates are reported in Table 10 in Online Resource 2. *Cort* appears overall as the most responsive gene, resulting the first in ranking for Amiodarone and Chlorpromazine (Table 2). The strong responsiveness of *Cort* is confirmed by the individual-experiment log(LOEC) estimates reported in Table 10 in Online Resource 2, where we notice that *Cort* appears as the most responsive gene also to Cyclosporine A. *Cort* is followed by *Nov*, ranking first for Diazepam and showing high responsiveness to Cyclosporine A, and possibly *Bdnf*. Note that all the fitted genes, with exception of *Hsp32* and *Gfap*, fall into the decreasing dose–response relationship scenario. This is in agreement with the role of *Hsp32* and *Gfap* which are known to be upregulated during cellular stress.

Normality of the residuals is assessed via the Shapiro–Wilk test. Table 11 in Online Resource 2 provides the resulting *p*-value for each fitted dataset. Values are generally above or equal to 5%, and all above 4%, indicating a fulfillment of the normality assumption.

Discussion

In this work, we have addressed model-based estimation of the LOEC by applying the Box–Cox transform both sides approach. We have additionally sought a summary of the results from three replicate experiments, as available in the experimental dataset. A model-based estimation provides significant advantages with respect to approaches where LOEC is identified by the first observed concentration level at which the response is significantly different from a control. The advantages arise from the fact that the whole dose–response relationship curve is inferred from the available data: as the LOEC estimate is obtained by inverse regression, it can correspond to any concentration level. Moreover, its uncertainty can be summarised in an appropriate confidence interval.

The simulation study has shown that the estimation process in this modeling setup has an overall tendency towards under-coverage, which becomes stronger when bootstrap rather than Wald-Delta method confidence intervals are considered. The observed behavior is probably to be linked with the challenges of nonlinear estimation combined with small sample sizes and a small number of groups for each drug–gene combination, and indeed less marked differences between bootstrap and Wald-Delta method confidence intervals coverage are observed for the 96-well plate design. In the real-data design, we have observed that enforcing parameter constraints does not significantly impact coverage properties, but may lead to a smaller number of converged runs. For the 96-well plate design, some advantages in estimation can be achieved by constrained optimisation, with small losses in terms of converged runs. The simulation study has also highlighted a potential multimodality in the target log-likelihood. As a practical advice, in scenarios comparable to those considered in the simulation study, Wald-Delta confidence intervals may represent a robust choice to interval estimation, which also comes at a lower computational cost; if bootstrap confidence intervals are sought, preference should be given to the semi-parametric approach. Adopting dispersed starting points is recommended, as it can reduce estimation error, and more generally can allow to better explore the log-likelihood target surface. Finally, the Box–Cox transformation may improve results when departures from normality/homoscedasticity are strong, but otherwise the increased length of the intervals may counterbalance difficulties in estimation, as for the Wald-Delta confidence interval approach. For the real-data design, estimation of the Box–Cox transformation parameter is more reliable through an ANOVA than a maximum likelihood estimation approach in decreasing dose–response relationships, but not for increasing ones. However, in the 96-well plate design, the ANOVA approach always results superior. A comparison of model residuals in the transformed scale based on both an ANOVA and a maximum-likelihood estimate of the Box–Cox parameter may provide further insight into which of the two methods estimates the transformation parameter more reliably and thus guide in the choice between the two approaches.

The data application, which provides the motivation for our research, shows that *Cort* is the overall most responsive gene, achieving the lowest LOEC estimates among all genes for three of the drugs considered; the analysis additionally demonstrates a good responsiveness of *Nov* and, to a weaker extent, *Bdnf*. BDNF is a neurotrophic factor which plays a crucial role in development and maintenance of neurons in the central nervous system, where it potentiates synaptic transmission. Modifications in brain, blood and cerebrospinal fluid levels of BDNF are associated with neurodegenerative and psychiatric diseases (Spulber et al. 2010; Mohammadi et al. 2018). A downregulation of BDNF mRNA in the rat brain is also observed after exposure to methylmercury (Andersson et al. 1997) and chronic administration of cyclosporine A (Chen et al. 2010). Furthermore, decreased circulating levels of BDNF are associated with alcohol-induced cognitive deficits (Silva-Peña et al. 2018). CORT is a neuropeptide expressed in distinct populations of

inhibitory neurons in the cerebral cortex and hippocampus (de Lecea et al. 1997) where it has anti-convulsant effects and controls sleep slow-wave activity (Hill et al. 2019). NOV is a member of the CCN family of proteins which are key players during organogenesis. However, although the central nervous system is a major site of NOV expression during brain development, its functions remain elusive (Le Dréau et al. 2009). It has recently been suggested that NOV plays a role in astrocyte activation and myelin regeneration (Le Dréau et al. 2009; Dombrowski et al. 2017). In this study, CORT, NOV and BDNF expression was even more sensitive to drug exposure than the four genes (GFAP, NFH, MBP and HSP32) we previously reported to be highly reliable markers of acute neurotoxicity (Zurich et al. 2013). This fact, together with the important described roles of BDNF, CORT and NOV in brain development and function suggest these three genes as good candidates for in vitro biomarkers of toxicity.

Acknowledgements The work of Marie-Gabrielle Zurich and Denise Tavel was supported by the European Union's Seventh Framework Collaborative Large-Scale Integrating Project Predict-IV under Grant agreement no. 202222. The authors wish to thank the anonymous reviewers for the valuable comments.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Andersson H, Lindqvist E, Olson L (1997) Downregulation of brain-derived neurotrophic factor mRNA in adult rat brain after acute administration of methylmercury. Mol Chem Neuropathol 31(3):225–233

Anon (2006) Background review document (BRD): validation of neutral red uptake test methods nih/in vitro cytotoxicity test methods for estimating acute oral systemic toxicity. https://ntp.niehs.nih.gov/iccvam/docs/acutetox_docs/brd_tmer/brdvol1_nov2006.pdf. Accessed 18 July 2019

Bal-Price AK, Hogberg HT, Buzanska L, Coecke S (2010) Relevance of in vitro neurotoxicity testing for regulatory requirements: challenges to be considered. Neurotoxicol Teratol 32(1):36–41

Bates DM, Watts DG (1988) Nonlinear regression analysis and its applications. Wiley, New York

Bickel PJ, Doksum KA (1981) An analysis of transformations revisited. J Am Stat Assoc 76(374):296–311

Box GE, Cox DR (1964) An analysis of transformations. J Roy Stat Soc Ser B (Methodol) 26:211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Carroll RJ, Ruppert D (1984) Power transformations when fitting theoretical models to data. J Am Stat Assoc 79(386):321–328

Carroll RJ, Ruppert D (1988) Transformation and weighting in regression, vol 30. CRC Press, Boca Raton

Chen CC, Hsu LW, Huang LT, Huang TL (2010) Chronic administration of cyclosporine a changes expression of BDNF and TrkB in rat hippocampus and midbrain. Neurochem Res 35(7):1098–1104

de Lecea L, del Rio JA, Criado JR, Alcántara S, Morales M, Danielson PE, Henriksen SJ, Soriano E, Sutcliffe JG (1997) Cortistatin is expressed in a distinct subset of cortical interneurons. J Neurosci 17(15):5868–5880

Dombrowski Y, O'Hagan T, Dittmer M, Penalva R, Mayoral S, Bankhead P, Fleville S, Eleftheriadis G, Zhao C, Naughton M, Hassan R, Moffat J, Falconer J, Boyd A, Hamilton P, Allen I, Kissenpfennig A, Moynagh PN, Evergren E, Fitzgerald D (2017) Regulatory T cells promote myelin regeneration in the central nervous system. Nat Neurosci 20:674–680

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton

Fitzmaurice GM, Lipsitz SR, Parzen M (2007) Approximate median regression via the Box–Cox transformation. Am Stat 61(3):233–238

Harrell FE (2015) Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, New York

Hill JL, Jimenez DV, Mai Y, Ren M, Hallock HL, Maynard KR, Chen HY, Hardy NF, Schloesser RJ, Maher BJ, Yang F, Martinowich K (2019) Cortistatin-expressing interneurons require TrkB signaling to suppress neural hyper-excitability. Brain Struct Funct 224(1):471–483

Hinkley DV, Runger G (1984) The analysis of transformed data. J Am Stat Assoc 79(386):302–309

Honegger P, Lenoir D, Favrod P (1979) Growth and differentiation of aggregating fetal brain cells in serumfree defined medium. Nature 282:305–8

Jensen SM, Kluxen FM, Ritz C (2019) A review of recent advances in benchmark dose methodology. Risk Anal. https://doi.org/10.1111/risa.13324

Jiang X, Kopp-Schneider A (2014) Summarizing EC_{50} estimates from multiple dose–response experiments: a comparison of a meta-analysis strategy to a mixed-effects model approach. Biom J 56(3):493–512

Jiang X, Kopp-Schneider A (2015) Statistical strategies for averaging EC_{50} from multiple dose–response experiments. Arch Toxicol 89(11):2119–2127

Latif AM, Gilmour SG (2015) Transform-both-sides nonlinear models for in vitro pharmacokinetic experiments. Stat Methods Med Res 24(3):306–324

Le Dréau G, Nicot A, Benard M, Thibout H, Vaudry D, Martinerie C, Laurent M (2009) NOV/CCN3 promotes maturation of cerebellar granule neuron precursors. Mol Cell Neurosci 43:60–71

Mohammadi A, Amooeian VG, Rashidi E (2018) Dysfunction in brain-derived neurotrophic factor signaling pathway and susceptibility to schizophrenia, Parkinson's and Alzheimer's diseases. Curr Gene Ther 18(1):45–63

Monnet-Tschudi F, Zurich MG, Pithon E, van Melle G, Honegger P (1995) Microglial responsiveness as a sensitive marker for trimethyltin (TMT) neurotoxicity. Brain Res 690(1):8–14

National Research Council (2007) Toxicity testing in the 21st century: a vision and a strategy. The National Academies Press, Washington, DC

Nottingham QJ, Birch JB (2000) A semiparametric approach to analysing doseresponse data. Stat Med 19(3):389–404

Ritz C, Streibig JC (2008) Nonlinear regression with R. Springer, Berlin

Ritz C, Van der Vliet L (2009) Handling nonnormality and variance heterogeneity for quantitative sublethal toxicity tests. Environ Toxicol Chem 28(9):2009–2017

Ritz C, Gerhard D, Hothorn LA (2013) A unified framework for benchmark dose estimation applied to mixed models and model averaging. Stat Biopharm Res 5(1):79–90

Ritz C, Baty F, Streibig JC, Gerhard D (2016) Dose-response analysis using R. PLoS One 10(12):1–13

Silva-Peña D, García-Marchena N, Alén F, Araos P, Rivera P, Vargas A, García-Fernández MI, Martín-Velasco AI, Villanúa MA, Castilla-Ortega E, Santín L, Pavón FJ, Serrano A, Rubio G, Rodríguez de Fonseca F, Suárez J (2018) Alcohol-induced cognitive deficits are associated with decreased circulating levels of the neurotrophin BDNF in humans and rats. Addict Biol. https://doi.org/10.1111/adb.12668

Spulber S, Rantamäki T, Nikkilä O, Castrén E, Weihe P, Grandjean P, Ceccatelli S (2010) Effects of maternal smoking and exposure to methylmercury on brain-derived neurotrophic factor concentrations in umbilical cord serum. Toxicol Sci 117(2):263–269

Wheeler MW, Shao K, Bailer AJ (2015) Quantile benchmark dose estimation for continuous endpoints. Environmetrics 26(5):363–372

Zurich MG, Monnet-Tschudi F, Costa LG, Schilter B, Honegger P (2004) Aggregating brain cell cultures for neurotoxicological studies. Humana Press, Totowa, pp 243–266

Zurich MG, Stanzel S, Kopp-Schneider A, Prieto P, Honegger P (2013) Evaluation of aggregating brain cell cultures for the detection of acute organ-specific toxicity. Toxicol In Vitro 27(4):1416–1424